ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization

Vadim Kantorov, Maxime Oquab, Minsu Cho, Ivan Laptev

▶ **To cite this version:**

Vadim Kantorov, Maxime Oquab, Minsu Cho, Ivan Laptev. ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization. ECCV 2016, Oct 2016, Amsterdam, Netherlands. pp.350 - 365, 10.1007/978-3-319-46454-1_22 . hal-01421772

HAL Id: hal-01421772
https://inria.hal.science/hal-01421772

Submitted on 22 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ContextLocNet: Context-Aware Deep Network Models for Weakly Supervised Localization

Vadim Kantorov, Maxime Oquab, Minsu Cho, and Ivan Laptev

WILLOW project team, Inria / ENS / CNRS, Paris, France
{vadim.kantorov,maxime.oquab,minsu.cho,ivan.laptev}@inria.fr

Abstract. We aim to localize objects in images using image-level supervision only. Previous approaches to this problem mainly focus on discriminative object regions and often fail to locate precise object boundaries. We address this problem by introducing two types of context-aware guidance models, *additive* and *contrastive* models, that leverage their surrounding context regions to improve localization. The additive model encourages the predicted object region to be supported by its surrounding context region. The contrastive model encourages the predicted object region to be outstanding from its surrounding context region. Our approach benefits from the recent success of convolutional neural networks for object recognition and extends Fast R-CNN to weakly supervised object localization. Extensive experimental evaluation on the PASCAL VOC 2007 and 2012 benchmarks shows that our context-aware approach significantly improves weakly supervised localization and detection.

Keywords: Object recognition, Object detection, Weakly supervised object localization, Context, Convolutional neural networks

1 Introduction

Weakly supervised object localization and learning (WSL) [1, 2] is the problem of localizing spatial extents of target objects and learning their representations from a dataset with only image-level labels. WSL is motivated by two fundamental issues of conventional object recognition. First, the strong supervision in terms of object bounding boxes or segmentation masks is difficult to obtain and prevents scaling-up object localization to thousands of object classes. Second, imprecise and ambiguous manual annotations can introduce subjective biases to the learning. Convolutional neural networks (CNN) [3, 4] have recently taken over the state of the art in many computer vision tasks. CNN-based methods for weakly supervised object localization have been explored in [5, 6]. Despite this progress, WSL remains a very challenging problem. The state-of-the-art performance of WSL on standard benchmarks [1, 2, 6] is considerably lower compared to the strongly supervised counterparts [7–9].

Strongly supervised detection methods often use contextual information from regions around the object or from the whole image [7, 9–13]: Indeed, visual context often provides useful information about which image regions are likely to

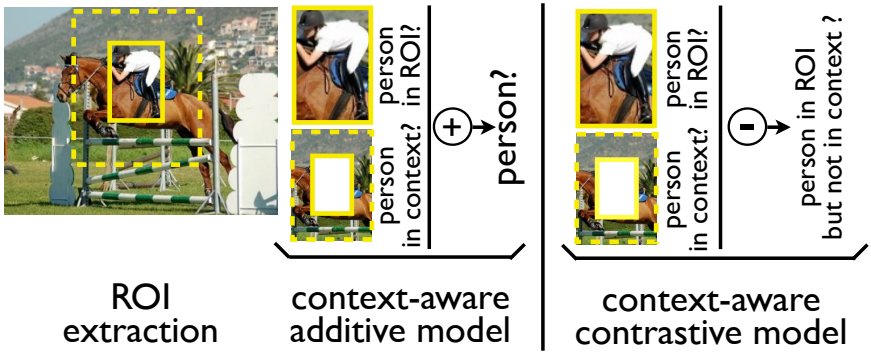


Fig. 1. Context-aware guidance for weakly supervised detection. Given extracted ROIs as localization candidates, our two basic context-aware models, *additive* and *contrastive* models, leverage their surrounding context regions to improve localization. The additive model relies on semantic consistency that aggregates class activations from ROI and context. The contrastive model relies on semantic contrast that computes difference of class activations between ROI and context. For details, see text. (Best viewed in color.)

be a target class according to object-background or object-object relations, e.g., a boat in the sea, a bird in the sky, a person on a horse, a table around a chair, etc. However, can a similar effect be achieved for object localization in a weakly supervised setting, where training data does not contain any supervisory information neither about object locations nor about context regions?

The main contribution of this paper is exploring the use of context as a supervisory guidance for WSL with CNNs. In a nutshell, we show that, even without strong supervision, visual context can guide localization in two ways: *additive* and *contrastive* guidances. As the conventional use of contextual information, the additive guidance enforces the predicted object region to be compatible with its surrounding context region. This can be encoded by maximizing the sum of a class score of a candidate region with that of its surrounding context. On the other hand, the contrastive guidance encourages the predicted object region to be outstanding from its surrounding context region. This can be encoded by maximizing the difference between a class score of the object region and that of the surrounding context. For example, let us consider a candidate box for a person and its surrounding region of context in Fig. 1. In additive guidance, appearance of a horse in the surrounding context helps us infer the surrounded region to contain a person. In contrast guidance, the absence of target-specific (person) features in its surrounding context helps separating the object region from its background.

In this work, we introduce two types of CNN architectures, *additive* and *contrastive* models, corresponding to the two contextual guidances. Building on the efficient region-of-interest (ROI) pooling architecture [8], the proposed models capture effective features among potential context regions to localize objects and

learn their representations. In practice we observe that our additive model prevents expansion of detections beyond object boundaries. On the other hand, the contrastive model prevents contraction of detections to small object parts. In experimental evaluation, we show that our models significantly outperform the baselines and demonstrate effectiveness of our models for WSL. The project webpage and the code is available at `http://www.di.ens.fr/willow/research/contextlocnet`.

2 Related Work

In both computer vision and machine learning, there has been a large body of recent research on WSL [1, 2, 5, 6, 14–24]. Such methods typically attempt to localize objects in the form of bounding boxes with visually consistent appearance in the training images, where multiple objects in different viewpoints and configurations appear in cluttered backgrounds. Most of existing approaches to WSL are formulated as or are closely related to multiple instance learning (MIL) [25], where each positive image has at least one true bounding box for a target class, and negative images contain false boxes only. They typically alternate between estimating a discriminative representation of the object and selecting an object box in positive images based on this representation. Since the task consists in a non-convex optimization problem, WSL has focused on robust initialization and effective regularization strategies.

Chum and Zisserman [14] initialize candidate boxes using discriminative visual words, and update localization by maximizing the average pairwise similarity across the positive images. Shi *et al.* [15] introduce the Latent Dirichlet Allocation (LDA) topic model for WSL, and Siva *et al.* [16] propose an effective negative mining approach combined with discriminative saliency measures. Deselaers *et al.* [17] instead initialize candidate boxes using the objectness method [26], and propose a CRF-based model that jointly localizes objects in positive training images. Song *et al.*formulate an initialization strategy for WSL as a discriminative submodular cover problem in a graph-based framework [19], and develop a negative mining technique to increase robustness against incorrectly localized boxes [20]. Bilen *et al.* [21] propose a relaxed version of MIL that softly labels object instances instead of choosing the highest scoring ones. In [22], they also propose a discriminative convex clustering algorithm to jointly learn a discriminative object model and enforce the similarity of the localized object regions. Wang *et al.* [1] propose an iterative latent semantic clustering algorithm based on latent Semantic Analysis (pLSA) that selects the most discriminative cluster for each class in terms of its classification performance. Cinbis *et al.* [2] extend a standard MIL approach and propose a multi-fold strategy that splits the training data to escape bad local optima.

As CNNs have turned out to be surprisingly effective in many vision tasks including classification and detection, recent state-of-the-art WSL approaches also build on CNN architectures [5, 6, 23, 24] or CNN features [1, 2]. Cinbis *et al.* [2] combine multi-fold multiple-instance learning with CNN features. Wang *et*

al. [1] develop a semantic clustering method on top of pretrained CNN features. While these methods produce promising results, they are not trained end-to-end. Oquab *et al.* [5] propose a CNN architecture with global max pooling on top of its final convolutional layer. Zhou *et al.* [24] apply global average pooling instead to encourage the network to cover the full extent of the object. Rather than directly providing the full extent of the object, however, these pooling-based approaches are limited to a position of a discriminative part or require a separate post-processing step to obtain the final localization. Jaderberg *et al.* [23] propose a CNN architecture with spatial transformer layers that automatically transform spatial feature maps to align objects to a common reference frame. Bilen *et al.* [6] modify a region-based CNN architecture [27] and propose a CNN with two streams, one focusing on recognition and the other one on localization, that performs simultaneously region selection and classification. Our work is related to these CNN-based MIL approaches that perform WSL by end-to-end training from image-level labels. In contrast to the above methods, however, we focus on a context-aware CNN architecture that exploits contextual relation between a candidate region and its surrounding regions.

While contextual information has been widely employed for object detection [7, 9, 11, 12, 28], the use of context has received relatively little attention in weakly supervised or unsupervised localization. Russakovsky *et al.* [29] and Cinbis *et al.* [2] use a background descriptor computed over features outside a candidate box, and demonstrate that background modelling can improve WSL as compared to foreground modelling only. Doersch *et al.* [30] align contextual regions of an object patch to gradually discovers a visual object cluster in their method of iterative region prediction and context alignment. Cho *et al.* [31, 32] propose a contrast-based contextual score for unsupervised object localization, which measures the contrast of matching scores between a candidate region and its surrounding candidate regions. Our context-aware CNN models are inspired by these previous approaches. We would like to emphasize that while the use of contextual information is not new in itself, we apply it to build a novel CNN architecture for WSL, that is, to the best of our knowledge, unique to our work. We believe that the simplicity of our basic models makes them extendable to a variety of weakly supervised computer vision tasks for more accurate localization and learning.

3 Context-Aware Weakly Supervised Network

In this section we describe our context-aware deep network for WSL. Our network consists of multiple CNN components, each of which builds on previous models [5, 6, 9, 27]. We begin by explaining first its overall architecture, and then detail our guidance models for WSL.

3.1 Overview

Following the intuition of Oquab *et al.* [5], our CNN-based approach to WSL learns a network from high-scoring object candidate regions within a classifi-

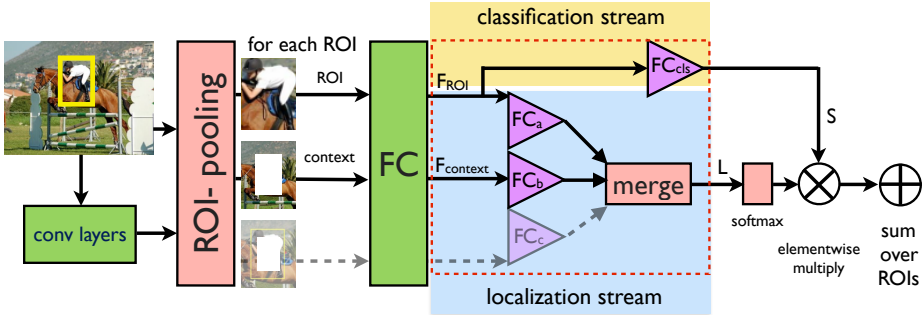


Fig. 2. Our context-aware architecture. Convolutional layers and FC layers (in green) correspond to the VGG-F architecture, pre-trained on ImageNet. The output of FC layers is passed through ReLu to the *classification* and *localization* streams. The classification stream takes features from ROIs, feeds them to a linear layer FC_{cls}, and outputs classification scores S_{ROI}. The localization stream takes features from ROIs and their context regions, processes them through our context-aware guidance models, and outputs localization scores L_{ROI}. The final output is a product of classification and localization scores for each ROI and object class. FC_{cls}, FC_a, FC_b, FC_c (in purple) are fully-connected linear layers trained from scratch. See text for more details.

cation training setup. In this approach, the visual consistency of classes within the dataset allows the network to localize and learn the underlying objects. The overall network architecture is described in Fig. 2.

Convolutional and ROI Pooling Layers. Our architecture has 5 convolutional layers, followed by a ROI pooling layer that extracts a set of feature maps, corresponding to the ROI (object proposal). The convolutional layers, as our base feature extractor, come from the VGG-F model [33]. Instead of max pooling typically used to process output of the convolutional layers in conventional CNNs for classification [4, 5], however, we follow the ROI pooling of Fast R-CNN [27], an efficient region-based CNN for object detection using object proposals [34]. This network first takes the entire image as input and applies a sequence of convolutional layers resulting in feature maps (256 feature maps with the effective stride of 16 pixels). The network then contains a ROI-pooling layer [35], where ROIs (object proposals) extract corresponding features from the final convolutional layer. Given a ROI on the image and the feature maps, the ROI-pooling module projects the ROI on the feature maps, pools corresponding features with a spatially adaptive grid, and then forwards them through subsequent fully-connected layers. This architecture allows us to share computations in convolutional layers for all ROIs in an input image. Following [6], in this work, we initialize network layers using the weights of ImageNet-pretrained VGG-F model [33], which is then fine-tuned in training.

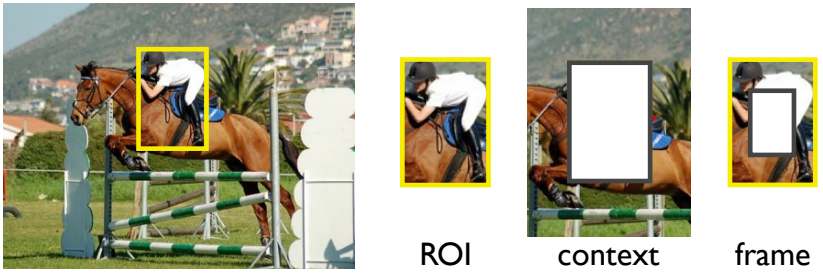
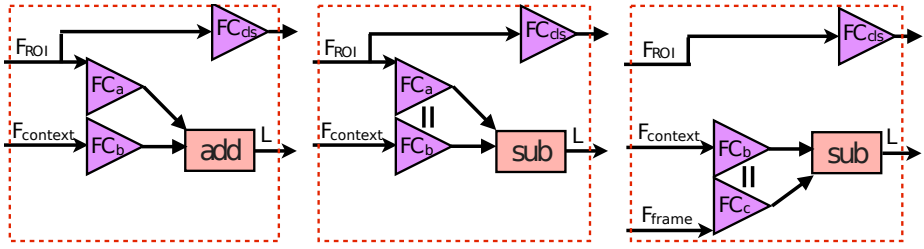


Fig. 3. Region pooling types for our guidance models: ROI pooling, context pooling, and frame pooling. For context and frame, the ratio between the side of the external rectangle and the internal rectangle is fixed as 1.8. Note that context and frame pooling types are designed to produce feature maps of the same shape, *i.e.*, frame-shaped feature maps with zeros in the center.

Feature Pooling for Context-Aware Guidance. For context-aware localization and learning, we extend the ROI pooling by introducing additional pooling types for each ROI, in a similar manner to Gidaris *et al.* [9]. As shown in Fig. 3, we define three types of pooling: ROI pooling, context pooling, and frame pooling. Given a ROI, *i.e.*, an object proposal [34], the *context* is defined as an outer region around the ROI, and the *frame* is an inner region ROI. Note that context pooling and frame pooling produce feature maps of the same shape, *i.e.*, central area of the outputs will have zero values. As will be explained in Sect. 3.3, this property is useful in our contrast model. The extracted feature maps are then independently processed by fully-connected layers (green FC layers in Fig. 2), that outputs a ROI feature vector, a context feature vector, and/or a frame feature vector. The models will be detailed in Sects. 3.2 and 3.3.

Two-Stream Network. To combine the guidance model components with classification, we employ the two-stream architecture of Bilen and Vedaldi [6], which branches a localization stream in parallel with a classification stream, and produces final classification scores by performing element-wise multiplication between them. In this two-stream strategy, the classification score of a ROI is reweighted with its corresponding softmaxed localization score. As illustrated in Fig. 2, the *classification stream* takes the feature vector F_{ROI} as input and feeds it to a linear layer FC_{cls}, that outputs a set of class scores S. Given C classes, processing K ROIs produces a matrix $S \in \mathbb{R}^{K \times C}$. The *localization stream* takes F_{ROI} and F_{context} as inputs, processes them through our guidance models, giving a matrix of localization scores $L \in \mathbb{R}^{K \times C}$. L is then fed to a softmax layer $[\sigma(L)]_{kc} = \frac{\exp(L_{kc})}{\sum_{k'=1}^{K} \exp(L_{k'c})}$ which normalizes the localization scores over the ROIs in the image. The final score for each ROI and class is obtained by element-wise multiplication of the corresponding scores S and $\sigma(L)$.

This procedure is done for each ROI and, as a final step, we sum all the ROI class scores to obtain the image class scores. During training, we use the hinge



(a) Additive model. (b) Contrastive model A. (c) Contrastive model S.

Fig. 4. Context-aware guidance models. The additive model takes outputs of ROI and context pooling, feeds them to independent fully-connected layers, and compute localization scores by adding their outputs. The contrastive models take outputs of ROI (or frame) and context pooling, feed them to a shared fully-connected layer (*i.e.*, two fully-connected layers with all parameter shared), and compute localization scores by subtracting the output of context from the other. For details, see the text.

loss function and train the model for multi-label image classification:

$$L(w) = \frac{1}{C \cdot N} \sum_{c=1}^{C} \sum_{i=1}^{N} \max(0, 1 - y_{ci} \cdot f_c(x_i; w)),$$

where $f_c(x; w)$ is the score of our model evaluated on input image x pararmeterized by w (all weights and biases) for a class c; $y_{ci} = 1$ if i'th image contains a ground truth object of class c, otherwise $y_{ci} = -1$. Note that the loss is normalized by the number of classes C and the number of examples N.

3.2 Additive Model

The additive model, inspired by the conventional use of contextual information [7, 9, 11, 12, 28], encourages the network to select a ROI that is semantically compatible with its context. Specifically, we introduce two fully-connected layers $\mathrm{FC}_{\mathrm{ROI}}$ and $\mathrm{FC}_{\mathrm{context}}$ as shown in Fig. 4 (a), and the localization score for each ROI is obtained by summing outputs of the layers. Note that compared to context-padding [7], this model separates a ROI and its context, and learns the adaptation layers $\mathrm{FC}_{\mathrm{ROI}}$ and $\mathrm{FC}_{\mathrm{context}}$ in different branches. This conjunction of separate branches allows us to learn context-aware activations for the ROI in an effective way.

Figure 5(top) illustrates the behavior of the $\mathrm{FC}_{\mathrm{ROI}}$ and $\mathrm{FC}_{\mathrm{context}}$ branches of the additive model trained on PASCAL VOC 2007. The scores of the target object (car) vary for different sizes of object proposals. We observe that the $\mathrm{FC}_{\mathrm{context}}$ branch discourages small detections on the interior of the object as well as large detections outside of object boundaries. $\mathrm{FC}_{\mathrm{context}}$ is, hence, complementary to $\mathrm{FC}_{\mathrm{ROI}}$ and can be expected to prevent detections outside of objects.

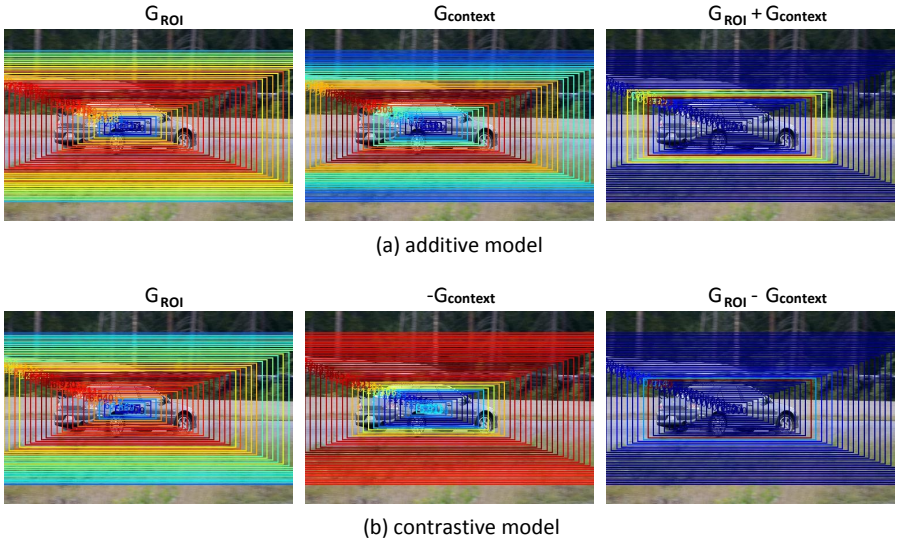


Fig. 5. Visualization of object scores produced by different branches of our models. The scores are computed for the *car* class for bounding boxes of different sizes centered on the target object. Red and blue colors correspond to high and low scores respectively. While the outputs of $\mathrm{FC}_{\mathrm{ROI}}$ branches for the additive and contrastive models are similar, the $\mathrm{FC}_{\mathrm{context}}$ branches, corresponding to feature pooling at object boundaries, have notably different behavior. The $\mathrm{FC}_{\mathrm{context}}$ branch of the additive model discourages detections outside of the object. The $\mathrm{FC}_{\mathrm{context}}$ branch of the contrastive model, discourages detections on the interior of the object. The combination of the $\mathrm{FC}_{\mathrm{ROI}}$ and $\mathrm{FC}_{\mathrm{context}}$ branches results in correct object localization for both models.

3.3 Contrastive Model

The contrastive model encourages the network to select a ROI that is outstanding from its context. This model is inspired by Cho *et al.*'s standout scoring for unsupervised object discovery [31], which measures the maximum contrast of matching scores between a rectangular box and its surrounding boxes. We adapt this idea of semantic contrast to our ROI-based CNN architecture. Specifically, we introduce two fully-connected layers $\mathrm{FC}_{\mathrm{ROI}}$ and $\mathrm{FC}_{\mathrm{context}}$ as shown in Fig. 4 (b), and the locacalization score for each ROI is obtained by subtracting the output activation of $\mathrm{FC}_{\mathrm{context}}$ from that of $\mathrm{FC}_{\mathrm{ROI}}$ for each ROI. Note that in order to make subtraction work properly, all weights of the layers $\mathrm{FC}_{\mathrm{ROI}}$ and $\mathrm{FC}_{\mathrm{context}}$ are shared for this model. Without sharing parameters, this model reduces to the additive model.

Figure 5(bottom) illustrates the behavior of $\mathrm{FC}_{\mathrm{ROI}}$ and $\mathrm{FC}_{\mathrm{context}}$ branches of the contrastive model. We denote by G_{ROI} and G_{context} the outputs of respective layers. The variation of scores for the car object class and different object proposals indicates low responses of $-G_{\mathrm{context}}$ on the interior of the object. The combination $G_{\mathrm{ROI}} - G_{\mathrm{context}}$ compensate each other resulting in correct localiza-

tion of object boundaries. We expect the contrastive model to prevent incorrect detections on the interior of the object.

One issue in this model is that in the localization stream the shared adaptation layers $\mathrm{FC}_{\mathrm{ROI}}$ and $\mathrm{FC}_{\mathrm{context}}$ need to process input feature maps of different shapes $\mathrm{F}_{\mathrm{ROI}}$ and $\mathrm{F}_{\mathrm{context}}$, *i.e.*, $\mathrm{FC}_{\mathrm{ROI}}$ processes features from a whole region (*ROI* in Fig. 3), whereas $\mathrm{FC}_{\mathrm{context}}$ processes features from a frame-shaped region (*context* in Fig. 3). We call this model the asymmetric contrastive model (*contrastive A*).

To remove this asymmetry in the localization stream, we replace ROI pooling with *frame* pooling (Fig. 3) that extracts a feature map from an internal rectangular frame of ROI. This allows the shared adaptation layers in the localization stream to process input feature maps of the same shape $\mathrm{F}_{\mathrm{frame}}$ and $\mathrm{F}_{\mathrm{context}}$. We call this model the symmetric contrastive model (*contrastive S*). Note that adaptation layer $\mathrm{FC}_{\mathrm{cls}}$ in the classification stream maintains the original ROI pooling regardless of modification in the localization stream. The advantage of this model will be verified in our experimental section.

4 Experimental Evaluation

4.1 Experimental Setup

Datasets and Evaluation Measures. We evaluate our method on PASCAL VOC 2007 dataset [36], which is a common benchmark in weakly supervised object detection. This dataset contains 2501 training images, 2510 validation images and 4952 test images, with bounding box annotations provided for 20 object classes. We use the standard trainval/test splits. We also evaluate our method on PASCAL VOC 2012 [37]. VOC 2012 contains the same object classes as VOC 2007 and is approximately twice larger in size for both splits.

For evaluation, two performance metrics are used: mAP and CorLoc. Detection mAP is evaluated using the standard intersection-over-union (IoU) criterion defined by [36]. Correct localization (CorLoc) [17] is a standard metric for measuring localization accuracy on a training set, where WSL usually provides one object localization per image for a target class. CorLoc is evaluated per-class, only on positive images for that class, and counts the percentage of images for which the highest-scoring candidate provided by the method overlaps (IoU > 0.5) with a ground truth box. We evaluate this mAP and CorLoc on the test and trainval splits respectively.

Implementation Details. ROIs for VOC 2007 are directly provided by the authors of the Selective Search proposal algorithm [34]. For VOC 2012, we use the Selective Search windows computed by Girshick *et al.* [27]. Our implementation is done using Torch [38], and we use the rectangular frame pooling based on the open-sourced code by Gidaris *et al.* [39, 40][1] which is itself based on Fast R-CNN [27] code. We use the pixel→features map coordinates transform for region

[1] `http://github.com/gidariss/locnet`

	Model	CorLoc	mAP
(a)	Cinbis *et al.* [2]	52.0	30.2
(b)	Wang *et al.* [1]	48.5	30.9
(c)	Wang *et al.*+ context [1]		31.6
(d)	WSDDN-SSW-S [6]		31.1
(e)	WSDNN-SSW-ENS [6]	54.2	33.3
(f)	WSDDN-SSW-S*	50.0	30.5
(g)	additive	52.8	33.3
(h)	contrastive A	50.2	32.2
(i)	contrastive S	**55.1**	**36.3**

Table 1. Comparison of our proposed models on PASCAL VOC 2007 with the state of the art, CorLoc (%) and detection mAP (%)

proposals from the public implementation of [35][2], with offset parameter set to zero (see the precise procedure in our code online[1]). All of our models, including our reproduction of WSDDN, use the same transform. We use the ratio between the side of the external rectangle and the internal rectangle fixed to 1.8.[3] Our pretrained network is the VGG-F model [33] ported to Torch using the loadcaffe package [41]. We train our networks using cuDNN [42] on an NVidia Titan X GPU. All layers are fine-tuned. Our training parameters are detailed below.

Parameters. For training, we use stochastic gradient descent (SGD) with momentum 0.9, dampening 0.0 on examples using a batch size of 1. In our experiments (both training and testing) we use all ROIs for an image provided by Selective Search [34] that have width and height larger than 20 pixels. The experiments are run for 30 epochs each. The learning rates are set to 10^{-5} for the first ten epochs, then lowered to 10^{-6} until the end of training. We also use jittering over scales. Images are rescaled randomly into one of the five following sizes: $800 \times 608, 656 \times 496, 544 \times 400, 960 \times 720, 1152 \times 864$. Random horizontal flipping is also applied.

At test time, the scores are evaluated on all scales and flips, then averaged. Detections are filtered to have a minimum score of 10^{-4} and then processed by non-maxima suppression with an overlap threshold of 0.4 prior to mAP calculation.

4.2 Results and Discussion

We first evaluate our method on the VOC 2007 benchmark and compare results to the recent methods for weakly-supervised object detecton [1, 6] in Table 1.

[2] `http://github.com/ShaoqingRen/SPP_net`

[3] This choice for the frame parameters follows [39, 40], and the ratio is kept same for both context and frame pooling types. We have experimented with different ratios, and observed that results of our method change marginally with increasing the ratio, and drop with decreasing the ratio.

Specifically, we compare to the WSDDN-SSW-S setup of [6] which, similar to our method, uses VGG-F as a base model and Selective Search Windows object proposals. For fair comparison we also compare results to our re-implementation of WSDDN-SSW-S (row (f) in Table 1). The original WSDDN-SSW-S employs an additional softmax in the classification stream and uses binary cross-entropy instead of hinge loss, but we found that these differences to have minor effect on the detection accuracy in our experiments (performance matches up to 1%, see rows (d) and (f)).

Our best model, contrastive S, reaches 36.3% mAP and outperforms previous WSL methods using selective search object proposals in rows (a)-(e) of Table 1. Class-specific CorLoc and AP results can be found in Tables 2 and 3, respectively.

Bilen *et al.* [6] experiment with alternative options in terms of EdgeBox object proposals, rescaling ROI pooling activations by EdgeBoxes objectness score, a new regularization term and model ensembling. When combined together, these additions improve result in [6] to 39.3%. Such improvements are orthogonal to our method and we believe our method will benefit from extensions proposed in [6]. We note that our single contrastive S model (36.3% mAP) outperforms the ensemble of multiple models using SSW in [6] (33.3% mAP).

Context Branch Helps. The additive model (row (g) in Table 1) improves localization (CorLoc) and detection (mAP) over those of the WSDDN-SSW-S* baseline (row (f)). We also applied a context-padding technique [7] to WSDDN-SSW-S* by enlarging ROI to include context (in the localization branch). Our additive model (mAP 33.3%) surpasses the context-padding model (mAP 30.9%). Contrastive A also improves localization and detection, but performs slightly worse than the additive model (Table 1, rows (g) and (h)). These results show that processing the context in a separate branch helps localization in the weakly supervised setup.

Contrastive Model with Frame Pooling. The basic contrastive model above, contrastive A (see Fig. 4), processes different shapes of feature maps (F_{ROI} and $F_{context}$) in the localization branch while sharing weights between FC_{ROI} and $FC_{context}$. To the contrary, contrastive S processes the same shape of feature maps (F_{frame} and $F_{context}$) in the localization branch. As shown in rows (h) and (i) of Table 1, contrastive S greatly improves CorLoc and mAP over contrastive A. Our hypothesis is that, since the weights are shared between the two layers in the the localization branch, these layers may perform better if they process the same shape of feature maps. Contrastive S obtains such a property by using frame pooling. This modification allows us to significantly outperform the baselines (rows (a) - (e) in Table 1). We believe that the model overfits less to the central pixels, achieving better performance. Per-class results are presented in Tables 2 and 3.

PASCAL VOC 2012 Results. The per-class localization results for the VOC 2012 benchmark using our contrastive model S are summarized in Table 4(de-

Model	aer	bik	brd	boa	btl	bus	car	cat	cha	cow	tbl	dog	hrs	mbk	prs	plt	shp	sfa	trn	tv	mAP
Cinbis *et al.* [2]	39.3	43.0	28.8	20.4	8.0	45.5	47.9	22.1	8.4	33.5	23.6	29.2	38.5	47.9	20.3	20.0	35.8	30.8	41.0	20.1	30.2
Wang *et al.* [1]	48.8	41.0	23.6	12.1	11.1	42.7	40.9	35.5	11.1	36.6	18.4	35.3	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	30.9
Wang *et al.*+context [1]	48.9	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	18.8	34.4	35.4	52.7	19.1	17.4	35.9	33.3	34.8	46.5	31.6
WSDDN-SSW-S*	49.8	50.5	30.1	**12.7**	11.4	54.2	49.2	20.4	1.5	31.2	27.9	18.6	32.2	49.7	**22.9**	15.9	25.6	27.4	38.1	41.3	30.5
additive	48.7	50.7	29.5	12.3	**14.1**	**56.5**	51.7	21.1	**4.0**	30.0	36.5	22.5	42.6	56.2	21.5	**17.5**	**29.5**	27.0	41.3	**52.3**	33.3
contrastive A	52.8	49.6	28.9	6.8	10.9	50.4	52.2	**35.0**	3.2	31.4	37.6	39.7	44.1	53.4	10.7	17.4	24.2	30.9	37.8	26.9	32.2
contrastive S	**57.1**	**52.0**	**31.5**	7.6	11.5	55.0	**53.1**	34.1	1.7	**33.1**	**49.2**	**42.0**	**47.3**	**56.6**	15.3	12.8	24.8	**48.9**	**44.4**	47.8	**36.3**

Table 2. Per-class comparison of our proposed models on VOC 2007 with the state of the art, detection AP (%)

Model	aer	bik	brd	boa	btl	bus	car	cat	cha	cow	tbl	dog	hrs	mbk	prs	plt	shp	sfa	trn	tv	avg
Conbis *et al.* [2]	65.3	55.0	52.4	48.3	18.2	66.4	77.8	35.6	26.5	67.0	46.9	48.4	70.5	69.1	35.2	35.2	69.6	43.4	64.6	43.7	52.0
Wang *et al.* [1]	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
WSDDN-SSW-S*	80.4	62.4	53.8	**28.2**	26.0	68.0	72.5	45.1	9.3	64.4	38.8	35.6	51.4	77.1	**37.6**	38.1	66.0	31.2	61.6	53.0	50.0
additive	78.8	66.7	52.9	25.0	**26.3**	68.0	73.6	44.8	**14.9**	62.3	45.2	46.3	61.6	82.3	35.3	39.6	**69.1**	30.9	62.0	**69.5**	52.8
contrastive A	78.8	62.7	51.1	20.2	21.8	68.5	71.6	**55.8**	10.3	**67.8**	46.8	53.7	62.2	82.3	26.0	**40.7**	55.7	33.6	55.5	39.4	50.2
contrastive S	**83.3**	**68.6**	**54.7**	23.4	18.3	**73.6**	**74.1**	54.1	8.6	65.1	**47.1**	**59.5**	**67.0**	**83.5**	35.3	39.9	67.0	**49.7**	**63.5**	65.2	**55.1**

Table 3. Per-class comparison of our proposed models on VOC 2007 with the state of the art, CorLoc (%)

tection AP) and Table 5(CorLoc). We are not aware of other weakly supervised localization methods reporting results on VOC 2012.

Observations. We have explored several other options and made the following observations. Training the additive model and the contrastive model in a joint manner (adding the outputs of individual models to compute the localization score that is further processed by softmax) have not improve results in our experiments. Following Gidaris *et al.* [40], we have tried adding other types of region pooling as input to the localization branch, however, this did not improve our results significantly. It is possible that different types of context pooling other than rectangular region pooling can provide improvements. We also found that sharing the weights or replacing the context pooling with the frame pooling in our additive model degrades the performance.

Qualitative Results. We illustrate examples of object detections by our method and WSDDN in Figure 6. We observe that our method tends to provide more accurate localization results for classes with localized discriminative parts. For example, for person and animal classes our method often finds the whole extent of the objects while previous methods tend to localize head regions. This is consistent with results in Table 2 where, for example, the dog class obtains the highest improvement by our contrastive S model when compared to WSDDN.

Our method still suffers from the second typical failure mode of weakly supervised methods, as shown in the two bottom rows of Figure 6, which is the multiple-object case: when many objects of the same class are encountered in close vicinity, they tend to be detected as a single object.

Model	aer	bik	brd	boa	btl	bus	car	cat	cha	cow	tbl	dog	hrs	mbk	prs	plt	shp	sfa	trn	tv	mAP
contrastive S	64.0	54.9	36.4	8.1	12.6	53.1	40.5	28.4	6.6	35.3	34.4	49.1	42.6	62.4	19.8	15.2	27.0	33.1	33.0	50.0	35.3

Table 4. Per-class comparison of the contrastive S model on VOC 2012 test set, AP (%)

Model	aer	bik	brd	boa	btl	bus	car	cat	cha	cow	tbl	dog	hrs	mbk	prs	plt	shp	sfa	trn	tv	Avg.
contrastive S	78.3	70.8	52.5	34.7	36.6	80.0	58.7	38.6	27.7	71.2	32.3	48.7	76.2	77.4	16.0	48.4	69.9	47.5	66.9	62.9	54.8

Table 5. Per-class comparison of the contrastive S model on VOC 2012 trainval set, CorLoc (%)

5 Conclusions

In this paper, we have presented context-aware deep network models for WSL. Building on recent improvements in region-based CNNs, we designed a novel localization architecture integrating the idea of contrast-based contextual guidance to the weakly-supervised object localization. We studied the localization component of a weakly-supervised detection network and proposed a subnetwork that effectively makes use of visual contextual information that helps refining the boundaries of detected objects. Our results show that the proposed semantic contrast is an effective cue for obtaining more accurate object boundaries. Qualitative results show that our method is less sensitive to the typical failure mode of WSL methods, such as shrinking to discriminative object parts. Our method has been validated on VOC 2007 and 2012 benchmarks demonstrating significant improvements over the baselines.

Given the prohibitive cost of large-scale exhaustive annotation, it is crucial to further develop methods for weakly-supervised visual learning. We believe the proposed approach is complementary to many previously explored ideas and could be combined with other techniques to foster further improvements.

Acknowledgments. We thank Hakan Bilen, Relja Arandjelović, and Soumith Chintala for fruitful discussion and help. This work was supported by the ERC grants VideoWorld and Activia, and the MSR-INRIA laboratory.

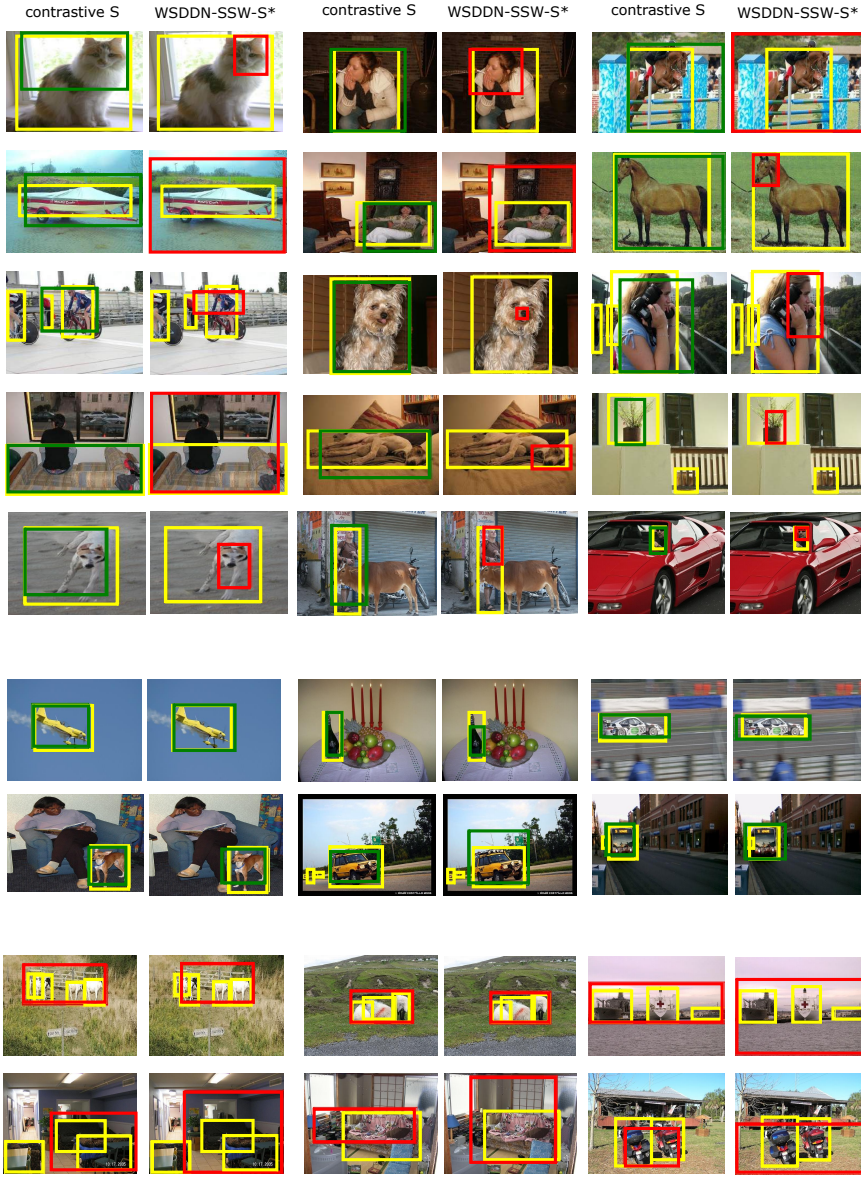


Fig. 6. The first five rows show localization examples where our method (contrastive S) outperforms WSDDN-SSW-S* baseline. Two next rows show examples where both methods succeed. The last two rows illustrate failure cases for both methods. Our method often suceeds in localizing correct object boundaries on examples where WSDNN-SSW-S* is locked to descriminative object parts such as heads of people and animals. Typical failure cases for both methods include images with multiple objects of the same class.

References

1. Wang, C., Ren, W., Huang, K., Tan, T.: Weakly supervised object localization with latent category learning. In: ECCV. Springer (2014) 431–445
2. Cinbis, R.G., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. arXiv preprint arXiv:1503.00949 (2015)
3. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural computation **1**(4) (1989) 541–551
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1097–1105
5. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: CVPR. (2015) 685–694
6. Bilen, H., Vedaldi, A.: Weakly supervised deep detection networks. In: CVPR. (2016)
7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. PAMI **38**(1) (2016) 142–158
8. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NIPS. (2015) 91–99
9. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware cnn model. In: ICCV. (2015) 1134–1142
10. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: ICCV, IEEE (2003) 273–280
11. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E., Belongie, S.: Objects in context. In: ICCV, IEEE (2007) 1–8
12. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI **32**(9) (2010) 1627–1645
13. Desai, C., Ramanan, D., Fowlkes, C.: Discriminative models for multi-class object layout. In: ICCV. (Sept 2009) 229–236
14. Chum, O., Zisserman, A.: An exemplar model for learning object classes. In: CVPR, IEEE (2007) 1–8
15. Shi, Z., Siva, P., Xiang, T., Mary, Q.: Transfer learning by ranking for weakly supervised object annotation. In: BMVC. Volume 2., Citeseer (2012) 5
16. Siva, P., Russell, C., Xiang, T.: In defence of negative mining for annotating weakly labelled data. In: ECCV. Springer (2012) 594–608
17. Deselaers, T., Alexe, B., Ferrari, V.: Weakly supervised localization and learning with generic knowledge. IJCV **100**(3) (2012) 275–293
18. Siva, P., Russell, C., Xiang, T., Agapito, L.: Looking beyond the image: Unsupervised learning for object saliency and detection. In: CVPR. (2013) 3238–3245
19. Song, H.O., Girshick, R., Jegelka, S., Mairal, J., Harchaoui, Z., Darrell, T.: On learning to localize objects with minimal supervision. arXiv preprint arXiv:1403.1024 (2014)
20. Song, H.O., Lee, Y.J., Jegelka, S., Darrell, T.: Weakly-supervised discovery of visual pattern configurations. In: NIPS. (2014)
21. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with posterior regularization. In: BMVC. (2014)
22. Bilen, H., Pedersoli, M., Tuytelaars, T.: Weakly supervised object detection with convex clustering. In: CVPR. (2015) 1081–1089
23. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: NIPS. (2015) 2008–2016

24. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. arXiv preprint arXiv:1512.04150 (2015)
25. Long, P.M., Tan, L.: Pac learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. Machine Learning **30**(1) (1998) 7–21
26. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. PAMI **34**(11) (2012) 2189–2202
27. Girshick, R.: Fast r-cnn. In: ICCV. (2015) 1440–1448
28. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in cognitive sciences **11**(12) (2007) 520–527
29. Russakovsky, O., Lin, Y., Yu, K., Fei-Fei, L.: Object-centric spatial pooling for image classification. In: ECCV. Springer (2012) 1–15
30. Doersch, C., Gupta, A., Efros, A.A.: Context as supervisory signal: Discovering objects with predictable context. In: ECCV. Springer (2014) 362–377
31. Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In: CVPR. (2015) 1201–1210
32. Kwak, S., Cho, M., Laptev, I., Ponce, J., Schmid, C.: Unsupervised object discovery and tracking in video collections. In: ICCV. (2015) 3173–3181
33. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. In: British Machine Vision Conference. (2014)
34. Uijlings, J.R., van de Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. IJCV **104**(2) (2013) 154–171
35. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. PAMI **37**(9) (2015) 1904–1916
36. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**(2) (2010) 303–338
37. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
38. Collobert, R., Kavukcuoglu, K., Farabet, C.: Torch7: A matlab-like environment for machine learning. In: BigLearn, NIPS Workshop. Number EPFL-CONF-192376 (2011)
39. Gidaris, S., Komodakis, N.: Locnet: Improving localization accuracy for object detection. arXiv preprint arXiv:1511.07763 (2015)
40. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware cnn model. In: ICCV. (2015) 1134–1142
41. Zagoruyko, S.: loadcaffe. `https://github.com/szagoruyko/loadcaffe` (2015)
42. Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B., Shelhamer, E.: cudnn: Efficient primitives for deep learning. arXiv preprint arXiv:1410.0759 (2014)